The LDCM Grid Prototype Overview

Jeff Lubelczyk
X45225
Jeffrey.T.Lubelczyk@nasa.gov

Agenda

- Prototype Background
- What is a Data Grid?
- LDCM Prototype Overview and Accomplishments

Prototype Background

- The objective of the LDCM Grid prototype (LGP) is to assess the applicability and effectiveness of a data grid to serve as the infrastructure for research scientists to generate virtual Landsat-like data products
- Grid technology serves as a key enabler in the creation of scientific Virtual Organizations, promotes a flexible and scalable infrastructure, facilitates the exchange of data, and maximizes the use of available resources
- A Grid infrastructure allows scientists at resource-poor sites access to remote resource-rich sites
  - Enables greater scientific research
  - Maximizes existing resources
  - Limits the expense of building new facilities

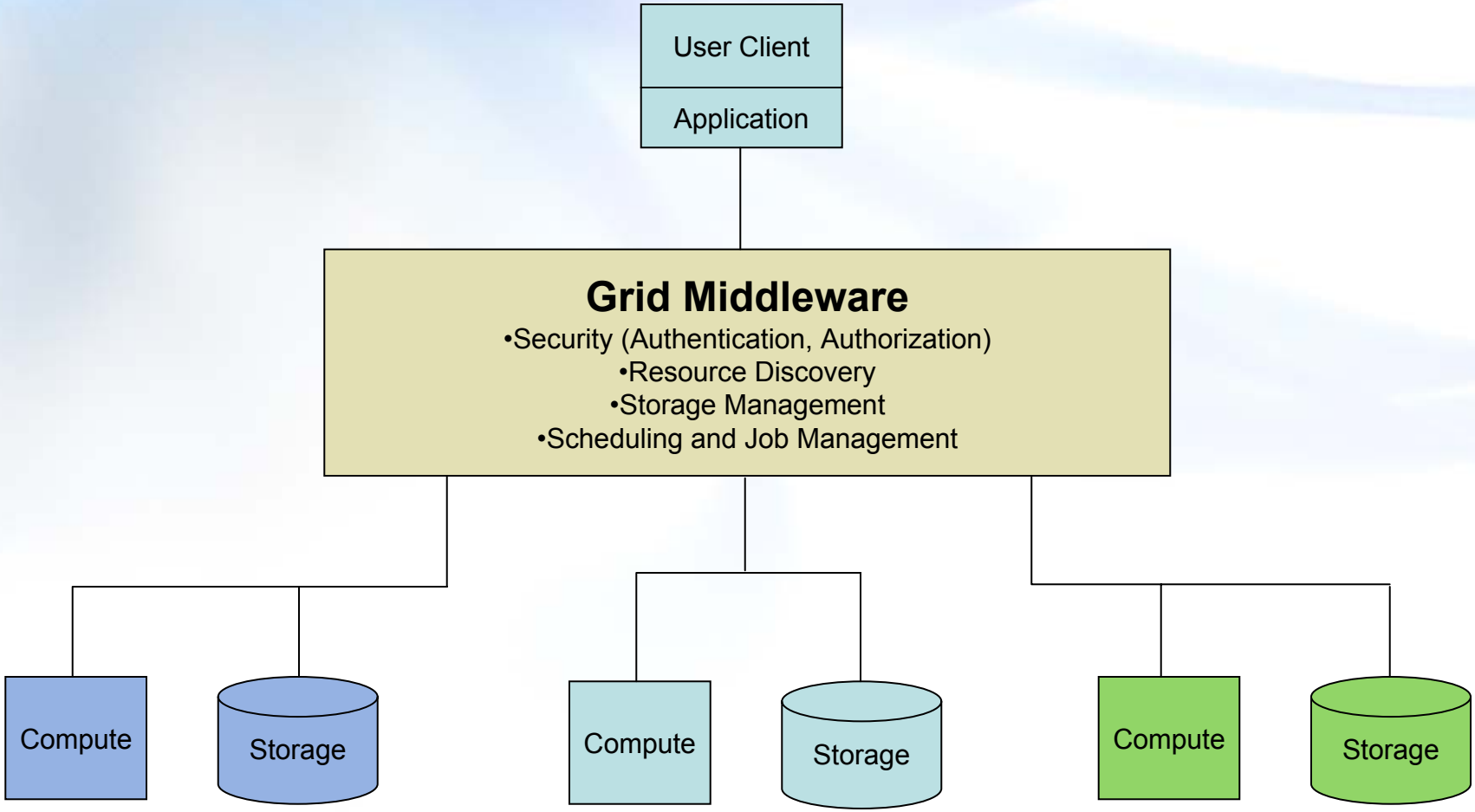
What is a data grid?

- In an article titled "Anatomy of the Grid," Ian Foster of Argonne National Labs suggests the following (2000):
  - "The sharing that we are concerned with is not primarily file exchange but rather direct access to computers, software, data, and other resources, as is required by a range of collaborative problem solving and resource-brokering strategies emerging in industry, science, and engineering. This sharing is, necessarily, highly controlled, with resource providers and consumers defining clearly and carefully just what is shared, who is allowed to share, and the conditions under which sharing occurs. A set of individuals and/or institutions defined by such sharing rules form what we call a *virtual organization*."
- He further suggests the following criteria:
  - Coordinates resources that are not subject to centralized control
  - Uses standard, open, general purpose protocols and interfaces
    - Otherwise, its an application specific system
  - Delivers nontrivial quality of service
    - Allows resources to be used in a coordinated fashion to deliver varying levels of service

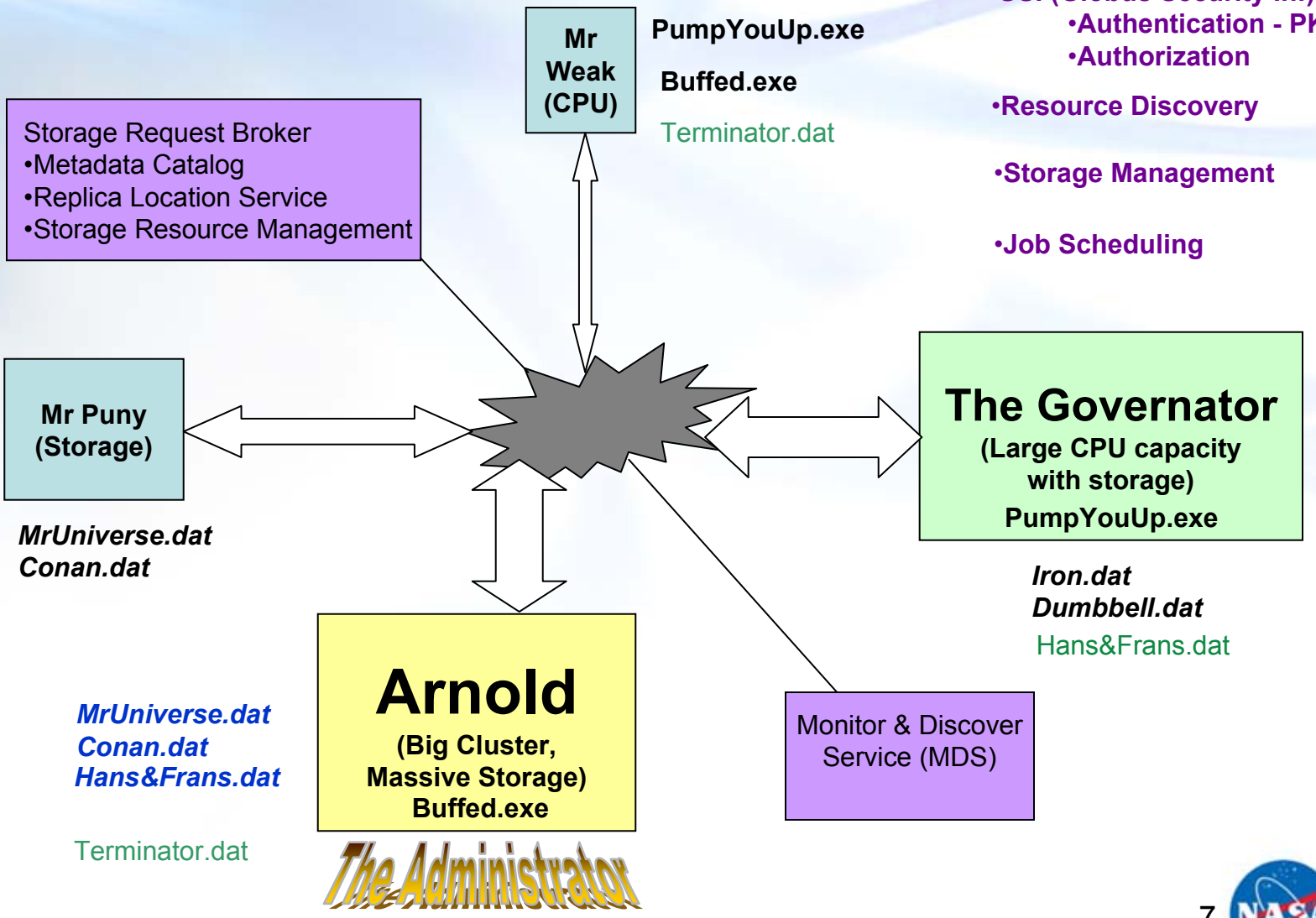
Data Grid

- **Grids provide a scaleable infrastructure**
  - Grid Software is the middleware
    - Provides a layer of abstraction
    - The underlying infrastructure is abstracted into defined APIs
    - A common package is the Globus Toolkit
  - Allows for dynamic collaboration of independently managed resources
    - Compute resources
    - Data Resources
    - Instruments and Sensors

Grid - A Layer of Abstraction



Let's look at what Grid can do for you



What the current data grid provides

- **Security**
  - Authentication (PKI)
  - Authorization
- **Resource Discovery**
  - Monitoring and Discovery Service (MDS) [LDAP like]
- **Storage Management and Brokering**
  - Metadata catalogs
  - Replica Location Service
    - Allows use of logical file names
    - Physical locations are hidden
    - Storage Resource Management
      - Retrieve data using physical file names
      - GridFTP
      - Data formats and subsetting
- **Job Scheduling and Resource Allocation**
  - GRAM (Globus Resource Allocation Manager) -- Provides a single common API for requesting and using remote system resources

Future Grid Capabilities

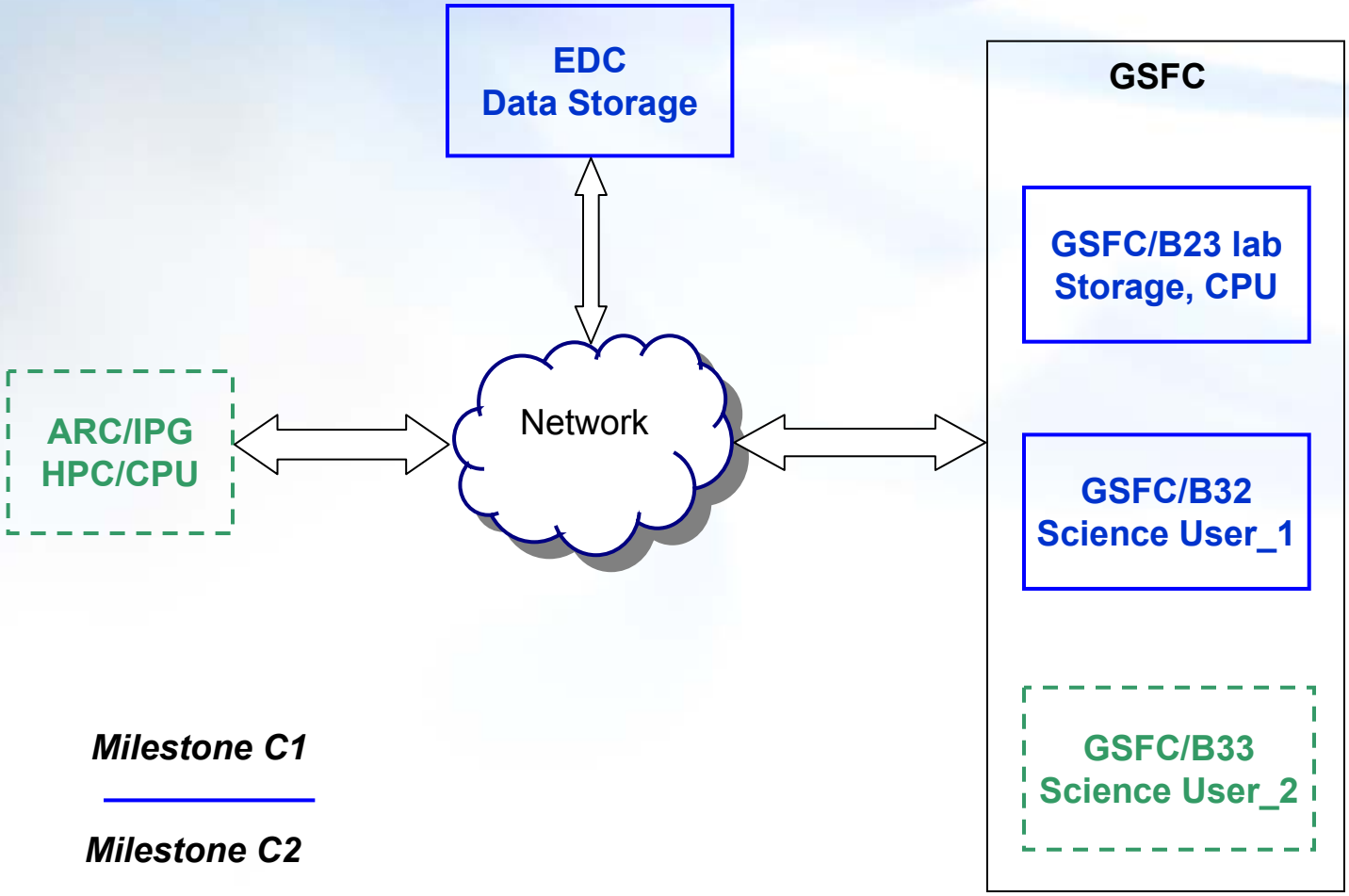
- **Intelligence**
  - Workflow Management
  - Automatic selection of resources to complete a given job or task
  - Intelligent Brokering -- Agent collaboration
- **Integration of grid and web services (WSRF)**
  - Standard proposed by the Globus Project (GTk 4.0)
- **The Global Grid Forum (GGF) serves as the international standards body for defining the Grid API/Framework**
  - Applications, Programming Models, and Environments
  - Architecture
  - Data
  - Grid Security
  - Information Systems and Performance
  - Peer to Peer
  - Scheduling Resource Management

LDCM Grid Prototype

Approach

- Capability 1
  - Provide and demonstrate a basic grid infrastructure that enables a user program to access remote heterogeneous instrument data at multiple GSFC labs and EDC.
- Capability 2
  - Enable the data fusion (blender) algorithm to obtain datasets, execute, and store the results on any resource within the Virtual Organization (GSFC labs, EDC, ARC IPG).

LGP Virtual Organization (C1 & C2)



Schedule

- Schedule
  - Prototype start (12/03)
  - Demo of Phase 1 grid infrastructure (6/04)
  - Demo of Phase 1 capability (12/04)
  - Demo of Phase 2 grid infrastructure (3/05)
  - Demo of Phase 2 capability (6/05)

Accomplishments

- **Established Partnerships**
  - USGS - Eros Data Center (EDC)
    - Agreed to supply limited storage and to facilitate the acquisition of identified data sets
  - Ames Research Center (ARC)
    - NASA leaders in the application of Grid technology
    - Information Power Grid (IPG) - Many large computing clusters. Supports a Highly Parallel Computing environment
    - Agreed to allow us to use their computing resources
  - Code 920- Data fusion algorithm development team
    - Working with them to obtain the data sets necessary to test the algorithm

Accomplishments (2)

- **LGP Team**
  - Code 580 supplying 2.5 FTEs (Civil Servants)
    - Lubelczyk, Weinstein, Ward, Kobler, Eng; McConaughy
  - Selected Grid Development contractor support
    - Held interviews with 4 potential contractor teams
    - Selected support contractors
      - SGT Corp
        - » Provide Grid development expertise and support
        - » 0.5 FTE now, ramping up to 0.8 FTE starting in June
      - Aerospace Corp
        - » Provide architectural consulting support
        - » Provide Grid installation, configuration, and administration support
        - » 0.2 FTE
  - Selected System Administration contractor support
    - QSS
      - Provides System Administration for Grid equipment at GSFC
      - 0.25 FTE

Accomplishments (3)

- **Configured GSI (IPG is our Certificate Authority)**
- **Set up a VO with 2 USGS servers and 2 sun machines in the 586 lab**
  - Configured grid mapfiles, FWs, and routers
  - Documented installation and config. procedures
- **Transferred sample ALI data file (250MB) from USGS using GridFTP**
  - Started at 17 minutes, now consistently at 28 seconds using GridFTP with 8 parallel streams
- **Finished test scenarios for installation checkout**
- **Finalized hardware architecture (Dell/Linux) and submitted PR**
- **Obtained and started using the UAH subsetting software on sample .hdf files (Capability 1)**
- **Assisting the MODIS team with Grid setup**
- **Participating in the CEOS Grid working group (International)**
  - Using a European grid tool called MapCenter
- **Set up a project Wiki**
- **Documenting Lessons Learned as we go**

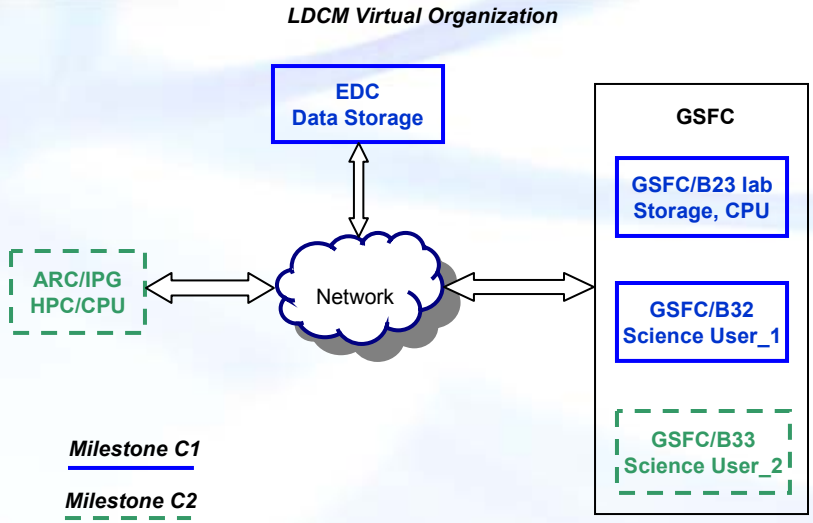
Back-up Slides

The LDCM Grid Prototype

POC: Jeff Lubelczyk, 586
Gail McConaughy, 586

Description and Objectives

- The objective of the LDCM ADG prototype is to assess the applicability and effectiveness of a data grid to serve as the infrastructure for research scientists to generate virtual Landsat-like data products.
- Grid technology serves as a key enabler in the creation of scientific Virtual Organizations, promotes a flexible and scalability infrastructure, facilitates the exchange of data, and maximizes the use of available resources



Approach

Phase 1: Provide and demonstrate a basic grid infrastructure that enables a simple data fusion algorithm to access remote heterogeneous instrument data at multiple GSFC labs and EDC.

Phase 2: Enable the data fusion algorithm to obtain datasets, execute, and store the results on any resource within the Virtual Organization (GSFC labs, EDC, ARC IPG).

Co-I's/Partners

EDC
NASA ARC/IPG
GSFC 920 Scientists

Science Themes

Virtual scientific data products
Remote instrument data access
Collaborative computing for the science community
Resource sharing and data discovery

Schedule and Deliverables

- Prototype start (12/03)
- Demo of Phase 1 grid infrastructure (6/04)
- Demo of Phase 1 capability (12/04)
- Demo of Phase 2 grid infrastructure (3/05)
- Demo of Phase 2 capability (6/05)

Application/Mission

Allow scientists at resource-poor sites access to remote resource-rich sites, enabling greater scientific research. Serve as a key enabler in the creation of scientific Virtual Organizations and by extension, facilities. Maximize utility of existing resources, limiting the expense of building new facilities.